Fall 12-3-2021

Computational Solutions to Exosomal microRNA Biomarker Detection in Pancreatic Cancer

Thuy T. An
University of Nebraska - Lincoln, thuy.an@huskers.unl.edu

COMPUTATIONAL SOLUTIONS TO EXOSOMAL MICRORNA BIOMARKER DETECTION IN PANCREATIC CANCER

by

Thuy Thi Hong An

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Juan Cui and Professor Jitender Deogun

Lincoln, Nebraska

December, 2021

COMPUTATIONAL SOLUTIONS TO EXOSOMAL MICRORNA BIOMARKER DETECTION IN PANCREATIC CANCER

Thuy Thi Hong An, MS

University of Nebraska, 2021

Adviser: Juan Cui, Jitender Deogun

Pancreatic cancer is the fourth leading cause of cancer death in the United States and the 5-year survival rate is only 5% to 10%. There are only a few non-specific symptoms associated with the early stage cancer, therefore most patients are diagnosed in a late stage. Due to the lack of effective treatments on pancreatic cancer and the fact that the early stage has a 39% 5-year survival rate, the biggest hope to control this disease is early detection. Therefore, discovery of effective and reliable non-invasive biomarkers for early detection of pancreatic cancer has been a major topic in the research field. Very recently, exosomal microRNAs have become promising candidates of diagnostic markers due to the facts that 1) such small non-coding RNA are stably present in the tissue and can get into blood circulation via exosome packaging which protects them from enzymatic degradation; 2) cancer cells, even at their early stages, may secrete up to tenfold more exosomes than normal cells and some disease-associated microRNAs can get into blood stream; 3) circulating exosomal microRNAs may carry early signals of cancers. With the goal to facilitate cancer detection, in this study, we have developed an integrated computational approach that leverages advanced genomics and bioinformatics to identify exosomal microRNAs that can be promising early detection biomarkers in pancreatic cancer. First, we have analyzed large-scale small

RNA microarray data collected from the GEO database to identify microRNA candidates that are differentially-expressed in cancer versus healthy control. Then we have explored several classification models to identify the most effective signature of microRNAs that can well differentiate cancer from normal based on the expression profiles. Through a series of comparisons and validations, the support vector machine-based classifier has achieved the best performance among all. A combination of five significantly-expressed microRNAs (hsa-miR-125a-3p, hsa-miR-6893-5p, hsa-miR-125b-1-3p, hsa-miR-6075, and hsa-miR-4294) achieved the best accuracy and AUC of 97.6% and 99.7%, respectively, which is highly promising. In order to explore the molecular determinants in microRNA secretion to further guide the non-invasive biomarker detection in the blood stream, in the second part of this project, we focused on sequence analysis of exosomal microRNAs for motif detection. Particularly, a graph-based motif-finding algorithm previously developed in our lab has been applied for this purpose. As a result, the motif [CUG][AU]G[UG] was found highly enriched in microRNAs associated with cancer exosomes. Knowing such properties is highly useful to guide a more targeted search in contrast to the profiling-based discovery where most of the detected microRNAs are highly abundant but likely disease irrelevant. In summary, our study has presented a new data-driven strategy that can potentially advance the biomedical research in biomarker discovery. Particularly, we have demonstrated that circulating exosomal microRNAs can be used as promising stable non-invasive biomarkers for early diagnosis of pancreatic cancer.

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my advisors, Dr. Juan Cui and Dr. Jitender Deogun, for their support and guidance during my study at the University of Nebraska - Lincoln (UNL).

Special thanks to Dr. Massimiliano Pierobon for being on my master thesis committee. I am extremely grateful for the financial support I've received from Dr. Juan Cui's grants and from the Department of Computer Science and Engineering (School of Computing) at UNL.

Special thanks to Haluk Dogan and Zeynep Hakguder for always answering tons of my questions and for their friendship.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The biggest challenge in treating cancer is that most cancers are not detected in the early stage, and due to evolution cancers advance in later stage when no effective treatment options are available. Therefore, early detection of cancer is the key to effective treatments for cancer patients [14]. Pancreatic cancer is an aggressive lethal disease in which cancer cells form in the tissues of the pancreas. The pancreas is located deep in the body and is surrounded by many crucial organs (Figure 1.1). The pancreas has two main functions, releasing enzymes that help food digestion and producing hormones that manage blood sugar levels. The digestive enzymes are made by exocrine pancreas cells and the hormones are made by neuroendocrine pancreas cells. Pancreatic cancer is classified into two types based on which part of the pancreas is affected [26]. Most pancreatic cancers arise in exocrine cells which make digestive enzymes. This cancer type is called pancreatic adenocarcinoma (PAAD) and is often found at an advanced stage [2]. The less common pancreas cancer type is formed in the hormone producing cells. This type of cancer is called pancreatic neuroendocrine tumors or pancreatic endocrine cancer and it usually has a

better prognosis. Pancreatic cancer is relatively rare, but it is the deadliest one. It is seldomly diagnosed early because there are few symptoms in early stages [14]. Pancreatic cancer does not cause symptoms until advanced stages when it has already spread to other organs. In addition, the symptoms of pancreatic cancer are similar to many other illnesses which make it hard to diagnose [9, 14].

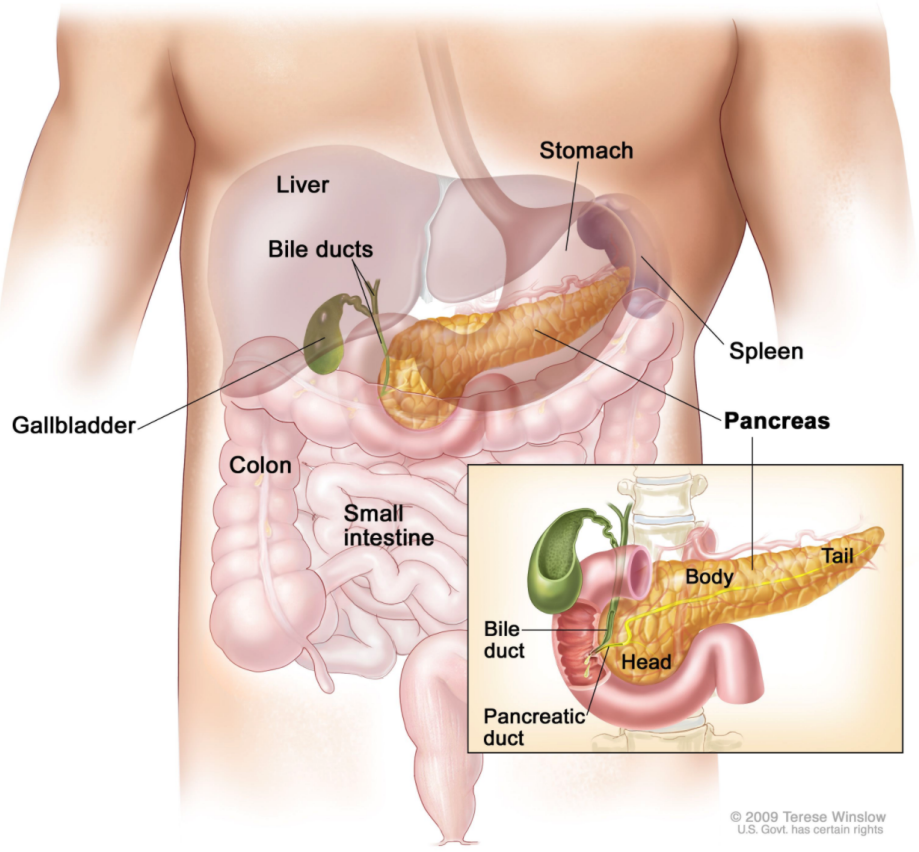

Figure 1.1: Anatomy of the pancreas [9]

1.2 Motivation

1.2.1 Challenges in early detection of pancreatic cancer

There are no biomarkers currently available for early diagnosis of pancreatic cancer. As a broad tumor biomarker, carbohydrate antigen 19-9 (CA19-9) is current used as a biomarker for pancreatic cancer in clinical practice [11]. However, CA19-9 is not specific to pancreatic cancer and it is not sensitive enough for early diagnosis of pancreatic cancer [26]. At the time of diagnosis, many patients have already reached advanced stage cancer [14]. In addition, pancreas is located deep in the body and surrounded by many organs, getting a tissue biopsy is a complex and risky task. Therefore, new biomarkers present in liquid biopsy, with high sensitivity and high specificity, are urgently needed for early detection of pancreatic cancer.

1.2.2 Roles of circulating exosomal microRNAs in early cancer detection

Exosomes can be isolated from multiple body fluids including blood plasma, serum, urine, breast milk, and saliva. Exosomes contain an abundance of small molecules (proteins, DNA, RNAs, etc.) which are pivotal in the cell-to-cell communication. It is believed that cancer cells secrete more exosomes than normal cells because they use exosomes as a mean of transportion for small molecules which needs to be spread to cultivate microenvironments for future metastasis [11]. Exosomes can regulate cell proliferation, invasion, metastasis, and participate in formation of chemoresistance in cancer treatment [14]. Many studies have indicated that proteins, microRNA, and mRNA enriched in exosomes are great candidates for biomarker cancer research regarding early

cancer diagnosis and determination of prognosis [3].

Among all the components in the exosomal cargos, exosomal microRNAs have been frequently reported to be the most promising biomarker candidates [26]. microRNAs play a key role in up and down regulation of pathways related to cancer development, tumorigenesis, metastasis, and resistance to various therapies. Many tumor-cell-released exosome-derived microRNAs are proangiogenic factors that promote neovascularization and metastasis. Exosomes are mean of transportation for transferring microRNAs from donor cells to target recipient cells, leading to reprogramming of the recipient cells. Exosomal microRNAs are more stable than free microRNAs because exosomes protect microRNAs from degradation by RNAase. Therefore, exosome is an enriched stable source of microRNAs in body fluids.

One of the challenges in developing microRNAs as biomarkers is the heterogenous nature of circulating microRNAs. microRNAs isolated from circulation have various forms including protein bound microRNAs and exosome associated microRNAs. The expression profile of microRNAs in exosomes is different from that in other microenvironment, suggesting that a strict sorting mechanism exists. Selective isolation of circulating microRNAs released from pancreatic cancer cells is a crucial step in the development of microRNAs as biomarkers. Since cancer cells release tenfold more exosomes than normal cells and certain microRNAs are selectively enriched in exosomes of cancer cells compared to exosomes of normal cells, circulating exosomal microRNAs are likely enriched in cancer-derived microRNAs [26]. Exosomes released by tumor cells likely contain microRNA biomarkers which can be used to predict

the presence of tumors in cancer patients. Opposite to tissue-based biopsy for cancer detection, liquid biopsies is less invasive, easier to collect, and less expensive. Therefore, circulating exosomal microRNAs are valuable biomarkers for early detection of cancer and biomarker research based on circulating exosomal microRNAs is an expending field [3].

Chapter 2

Background and Related Work

2.1 Background about microRNAs and exosomes

2.1.1 microRNA formation and function

microRNAs are small single-stranded, non-coding RNA molecules with an average of 22 nucleotides in length [16, 20]. There are four steps in microRNA formation (Figure 2.1). In the first step, microRNA gene is transcribed to generate a primary microRNA. Then primary microRNA is cleaved to form a precursor microRNA which is then exported into the cytoplasm. In the cytoplasm, precursor microRNA is undergone further cleavage to remove the stem loop and form a short double-stranded microRNA molecule called a duplex. Finally, the duplex unwinds to release one strand. The remaining strand combines with some proteins and forms a RNA-induced Silencing Complex (RISC). The RISC binds target mRNA to prevent protein production by one of two distinct gene silencing mechanisms. In the first mechanism, RISC just simply cuts the mRNA which will be further destroyed by the cell. The other mechanism is the translation repression in which the RISC prevents ribosome subunits from binding, thus, prevents translation from happening. Either way, mature microRNAs can regulate gene expression levels and control many pro-

cesses in the cells.

Cancer is a type of genetic disease in which abnormal cells grow uncontrollably by disregarding the normal rules of cell division. The failure in gene regulation involving in cell proliferation, differentiation and apoptosis is associated with cancer initiation and progression [16]. microRNAs play a crucial role in regulating genes that control these processes. Therefore, dysregulation of microRNA expression profiles can lead to cancer development.

2.1.2 Exosome formation and function

Exosomes are nanometer-sized (30-100 nm in diameter) membrane-derived lipid bilayer vesicles that contain constituents (proteins, lipids, and nucleic acids like DNA, mRNA, microRNA, etc.) of the cell that secrete them (Figure 2.2). Exosomes are the smallest extracellular vesicles.

Exosome formation is divided into five different stages (Figure 2.3): (A) Exosome is derived from early endosome formed from plasma membrane. (B) Early endosome becomes late endosomes. (C) Then forms early multivesicular bodies. (D) Late multivesicular bodies. (E) Late multivesicular bodies can either get degraded by lysosomes or fuse with the membrane to release exosomes [3]. Released exosomes can be taken by surrounding recipient cells (via fusion or receptor-ligand interaction) or travel through biological fluids (blood, urine, saliva) and later taken up by distant cells. Exosomes can be isolated from multiple body fluids such as blood plasma, serum, urine, breast milk, and saliva [14]. Molecules being packed in exosomes are not random selected. There are motif sequences that determine if some molecules are being packed or not.

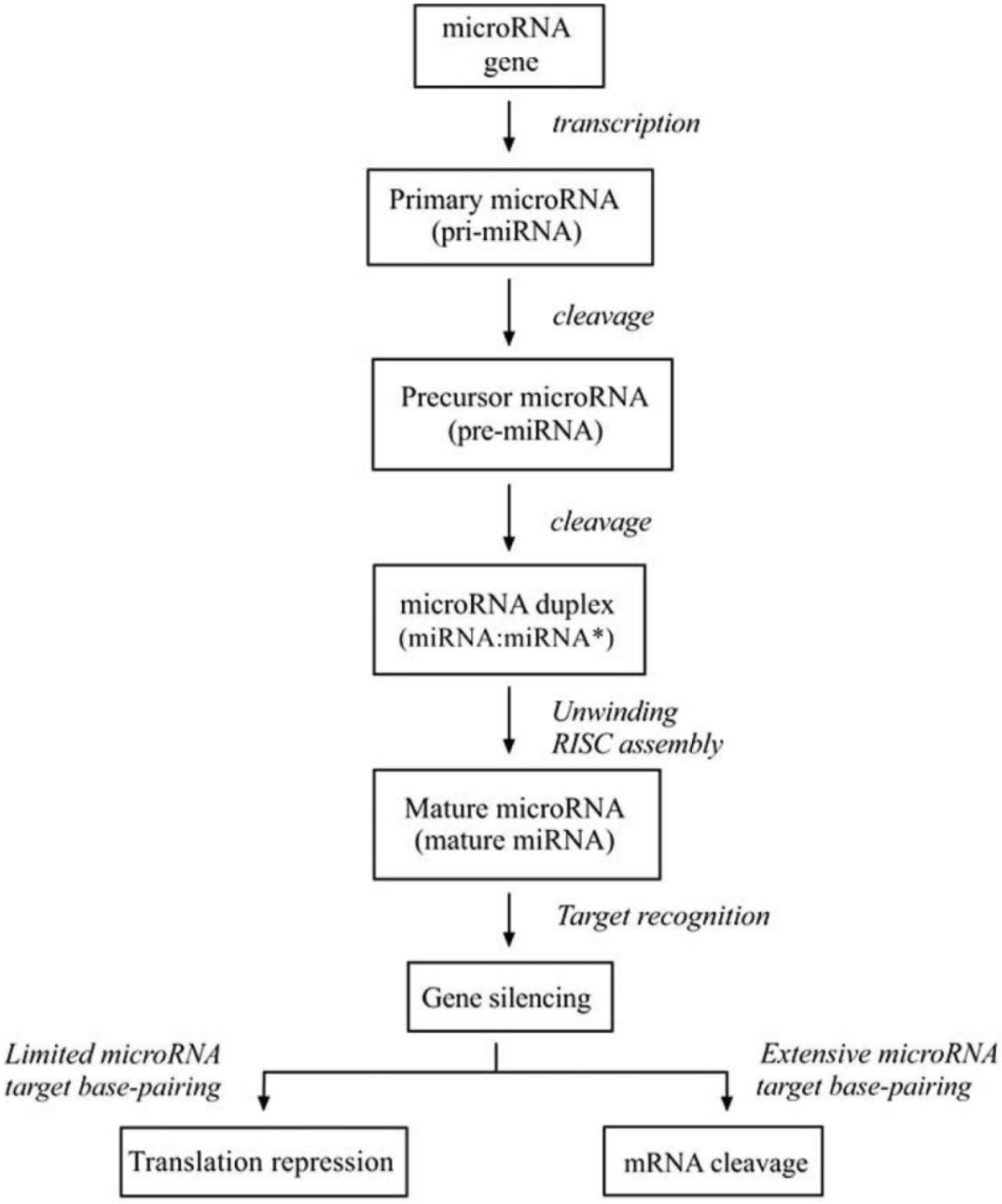

Figure 2.1: microRNA formation and function [16]

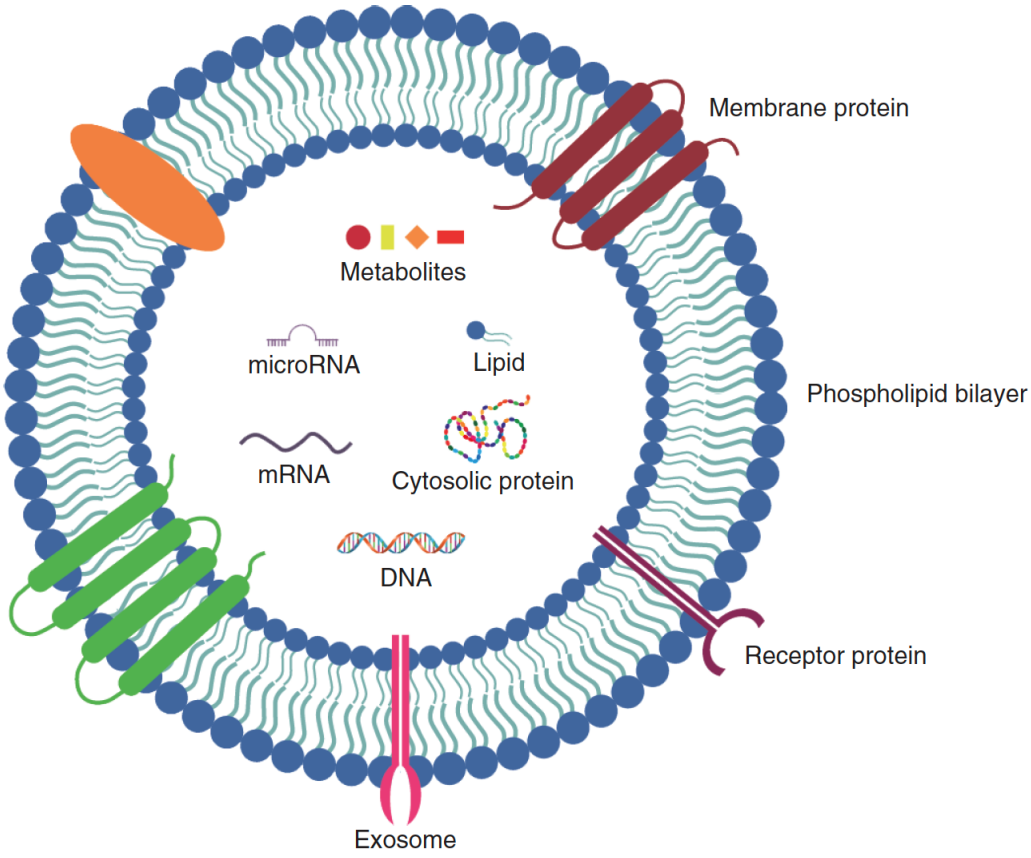

Figure 2.2: The structure and content of exosome [3]

Exosomes are important in intercellular communication because of their ability in transferring information from donor cells to recipient cells, leading to changes in the recipient cells. Cancer cells produce about tenfold more exosomes than normal healthy cells. Exosomes are extremely stable, and the contents packed in exosomes are valuable resources for biomarker research for early detection of cancer [14, 26, 3].

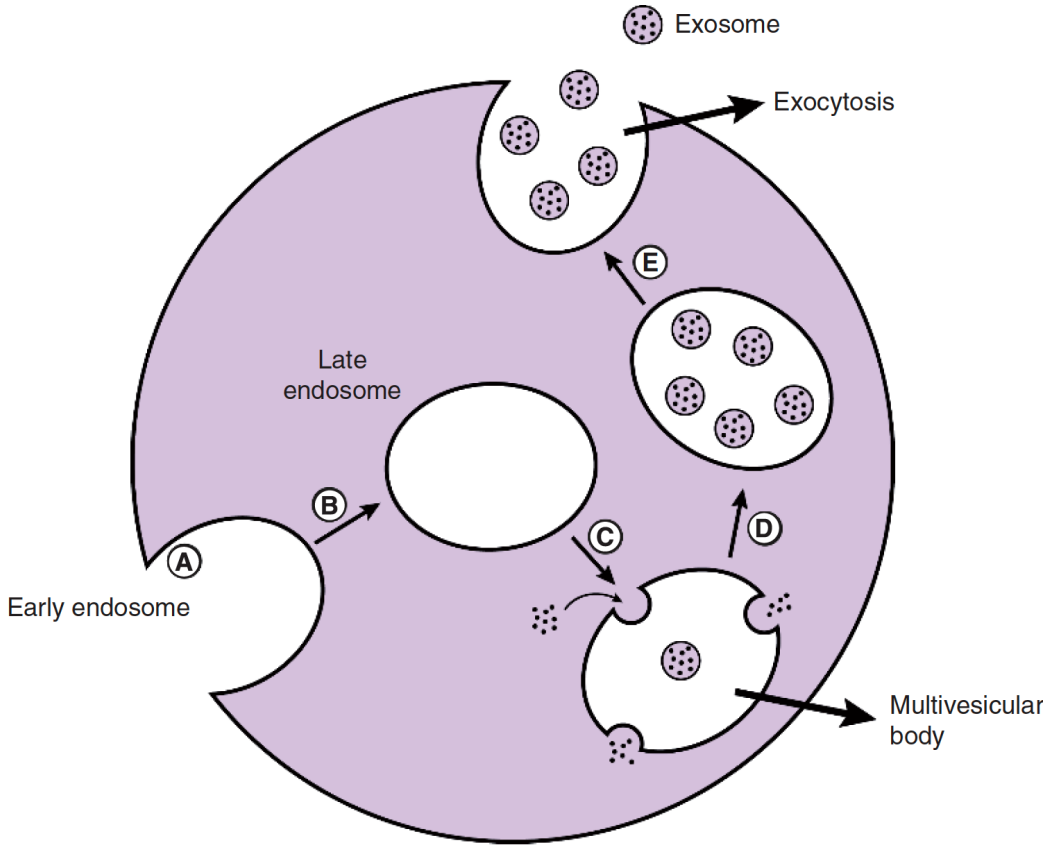

Figure 2.3: The formation and releasing of exosome [3]

2.2 Literature review on exosomal microRNAs as biomarkers for pancreatic cancer

The difference of microRNA expression between pancreatic cancer cells and normal healthy cells can be used to identify biomarkers for diagnostic and prognostic purposes. Expression of several microRNAs was elevated in the circulation of pancreatic cancer patients and several exosomal microRNAs have been reported as potential biomarkers [26]. The expression profiles of miR-17-5p and miR-21 were greatly enhanced in pancreatic cancer patients compared to patients with benign pancreatic diseases and healthy controls [14, 26]. Using the NGS technology and qRT-PCR, miR-10b, miR-550, miR-196a, miR-1246,

and miR-451a were determined to be present at higher levels in exosomes extracted from pancreatic cancer cells. Serum exosomal microRNAs, including miR-1246, miR-4644, miR-3976 and miR-4306, were significantly upregulated in patients with pancreatic cancer compared to patients with benign pancreatic disorders [26]. Table 2.1 summarized microRNAs which were identified as biomarkers in pancreatic cancer patients. These results suggest that some exosomal microRNAs can be used as potential liquid biopsy biomarkers for early diagnosis of pancreatic cancer [14].

Biomarkers	**Samples**	**Comments**	**Refs**
miR-2, miR-155	27 PDACs and 8 CPs	These exosomal microRNAs isolated from pancreatic juice were shown as promising biomarkers for pancreatic cancer	[19]
miR-1246, miR-4644	12 PCs (6 males, 6 females), 13 healthy participants	miR-1246 and miR-4644 in salivary exosomes could be useful biomarkers for identification of patients with pancreatobiliary tract cancer	[17]

miR-4525, miR-451a, miR-21	55 PDACs Compared expression levels of exosomal microRNAs in portal vein blood (PVB) and peripheral blood (PB)	Exosomal miR-4525, miR-451a and miR-21 levels were upregulated in PVB, thus, higher than those in the PB	[22]
miR-17-5p, miR-21	22 PCs, 6 benign pancreatic tumors, 7 ampullary carcinomas, 6 CPs, 8 healthy controls	Diagnostic biomarker for dividing PC and non-PC	[21, 14, 1]
miR-10b	3 PDACs, 3 CPs, 3 healthy controls	Diagnostic biomarker for PDAC comparing with CP and normal control	[14, 24]
miR-196a, miR-1246	15 PDACs (Stage I-IIA), 15 healthy controls	Diagnostic biomarker for dividing PDAC and non-PC. Plasma exosome miR-196a and miR-1246 levels were significantly elevated in pancreatic cancer patients as compared to healthy subjects	[28, 14, 23, 27]

miR-1246, miR-4644, miR-3976, miR-4306	20 healthy donors, 131 pancreatic cancer, 25 CP, 22 benign pancreatic tumors, 12 non pancreatic cancer	These biomarkers were significantly upregulated in patients with pancreatic cancer when compared to patients with benign pancreatic disorders. The authors suggested to use a combination of protein and microRNA biomarkers.	[18]
miR-451a	7 PDACs with stage I, 43 PDACs with stage II, 20 healthy controls	Minimally invasive biomarker for the prediction of recurrence and prognosis of pancreatic cancer	[14, 23]
miR-10b, miR-21, miR-30c, miR-181a	29 PDACs, 11 CPs	High expression of miR-10b, miR-21, miR-30c, and miR-181a and low expression of miR-let7a is a better indicator of pancreatic cancer than the exosomal glypican-1 levels	[13]
miR-191, miR-21, miR-451a	32 PCs, 29 IPMNs, 22 controls	The level of these microRNAs was significantly elevated in patients with pancreatic cancer and intraductal papillary mucinous neoplasm as compared to benign controls	[6]

A list of 39 pancreatic cancer biomarkers	63 PCs, 63 control subjects	Identify biomarkers for pancreatic cancer and build a prediction model	[15]
mir-744-5p, mir-409-3p, mir-128-3p	24 PCs, 21 healthy controls, 10 biliary tract cancer patients	Distinguishable patterns between cancer patients and healthy controls. However, they were unable to distinguish pancreatic cancer from biliary tract cancer	[10]

Table 2.1: microRNAs as biomarkers for early diagnosis of pancreatic cancer. PC stands for pancreatic cancer; PDAC stands for pancreatic ductal adenocarcinoma; CP stands for chronic pancreatitis; IPMN stands for intraductal papillary mucinous neoplasm.

The above studies suggested that exosomal microRNAs have great promise as biomarkers in the diagnosis and prognosis of pancreatic cancer. However, most studies have been conducted with a small number of samples and the results vary between studies. Each study suggests different circulating microRNAs as biomarkers for pancreatic cancer. Research projects at larger scales and better designs are needed to better identify potential biomarkers for cancer detection and improve their reproducibility. More efforts should be given to analyze pancreatic cancer exosomal microRNA signatures.

Chapter 3

Research Questions

Our goal is to develop an approach for identifying promising non-invasive exosomal microRNA biomarkers for early diagnosis and prognosis of pancreatic cancer.

3.1 Research question 1

What exosomal microRNAs can be used as potential non-invasive biomarkers for the detection of pancreatic cancer?

3.2 Research question 2

What are the motif sequences in exosomal microRNAs that make these microRNAs sorted into exosomes?

Chapter 4

Methodologies

4.1 Data Collection

4.1.1 Dataset 1: Microarray expression data on pancreatic cancer vs normal

We downloaded the microRNA expression profile of Series GSE59856 from the NCBI Gene Expression Omnibus (GEO) database using R script. This publicly available microarray data consist of 100 pancreatic cancer patients (pancreatic ductal adenocarcinoma) and 150 healthy controls (Figure 4.1). As described in the NCBI GEO Series GSE59856, total RNAs were isolated from serum samples using 3D-gene RNA extraction reagent. Comprehensive microRNA expression profiles were captured using a 3D-Gene microRNA labeling kit, a highly sensitive microarray. The downloaded data contain 103 missing values (NaN, Not a Number). We replaced each NaN with the average of five values just before it. The data table contained accession number (e.g., MIMAT0000062) instead of the name of microRNA (e.g., hsa-let-7a-5p). To make it easier to understand, we used Pandas dataframe in Jupyter Notebook to replace all accession numbers by their corresponding microRNA names before moving forward with further data analysis.

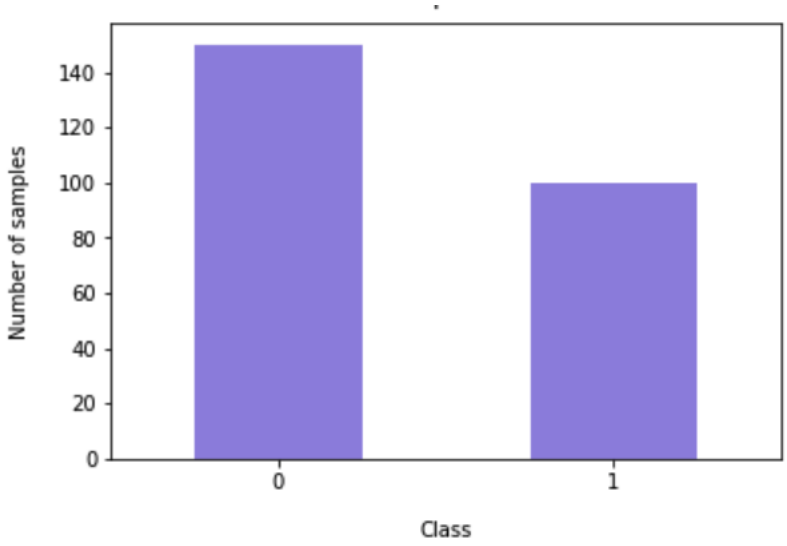

Figure 4.1: Number of samples in GSE59856 (pancreatic cancer: "1" and healthy control: "0")

4.1.2 Dataset 2: Exosomal miRNA sequences from ExoCarta database

We downloaded exosomal microRNA data from ExoCarta, an exosome database which provide with the contents that were identified in exosomes in multiple organisms. We selected only exosomal microRNAs which were found in human (Homo sapiens). We used fastaselect.pl ([4]) to extract the sequences of these microRNAs from the miRBase database v22 ([12]). As a result, 880 sequences were obtained for human exosomal microRNAs in the ExoCarta database.

4.2 Data analysis for Series GSE59856

4.2.1 Visualization of sample distributions in 2-D and 3-D

In this microRNA dataset, we have a total of 250 samples (150 healthy controls and 100 cancer samples) and 2555 features (microRNAs). The number of samples are much smaller than the number of features. This is a common big

p little n problem in microRNA datasets. We cannot see sample distribution in 2555 dimensions. Therefore, we performed Principal Component Analysis (PCA) for feature extraction so that we can see sample distributions in two- and three-dimensions. Normal healthy samples were labeled as "0" and pancreatic cancer samples were labeled as "1".

4.2.2 Building Machine Learning Models for Cancer Classification

In this research, machine learning models for binary classification of the pancreatic cancer versus healthy control were built and evaluated using Scikit-learn, a software machine learning library for Python. The following algorithms were chosen to be examined: SVM, decision tree classifier, linear discriminant analysis, quadratic discriminant analysis, and Gaussian naive bayes. The SVM algorithm has been widely applied in the biological sciences as a means of classifier for molecular data [29]. Decision tree classifier is one of the common machine learning methods to address the challenging task, large dimensionality and very small number of samples, of gene expression data classification [7]. Linear discriminant analysis is among the most effective procedures in the domain of high-dimensional prediction [8]. Quadratic discriminant analysis is also capable of handling with high dimensional data [2]. Gaussian naive bayes is naturally immune to the curse of dimensionality [25]. Data were splited into training samples and testing samples with a test size of 33%. The performance of the above models was evaluated using cross validation and some standardized evaluation metrics such as recall, precision, and F1 score.

4.2.3 Differential gene expression analysis using Limma

For microarray data Series GSE59856, we used Limma (R script, Bioconductor) to identify the microRNAs which were expressed differently in the exosomes of pancreatic patients and healthy donors. Limma stands for linear models for microarray data since it was historically designed to analyze microarray data. Quantile normalization was conducted to normalize the signals across the different microarrays tested. The level of differential gene expression was ranked based on B-statistics which was the same order with t-statistics in this case.

4.2.4 SVM-based models for cancer classification

Even though we have 2555 microRNAs in the dataset, not all microRNAs are equally important in distinguishing cancer samples from normal samples. Strongly differentially expressed microRNAs are likely the best biomarkers to separate these two groups. Here we built a SVM model for binary classification of cancer patients from healthy controls.

In order to see how many top differently express microRNAs are needed to differentiate pancreatic cancer patients from healthy controls, full microRNA profiles, top 20 differently expressed microRNAs, top 15 differently expressed microRNAs, top 10 differently expressed microRNAs, top 5 differently expressed microRNAs, and top 2 differently expressed microRNAs were used as input for the SVM model. We compared the performance of the SVM model when different numbers of microRNAs were used as inputs by using some standardized evaluation metrics such as precision, recall, F1 score, and confusion

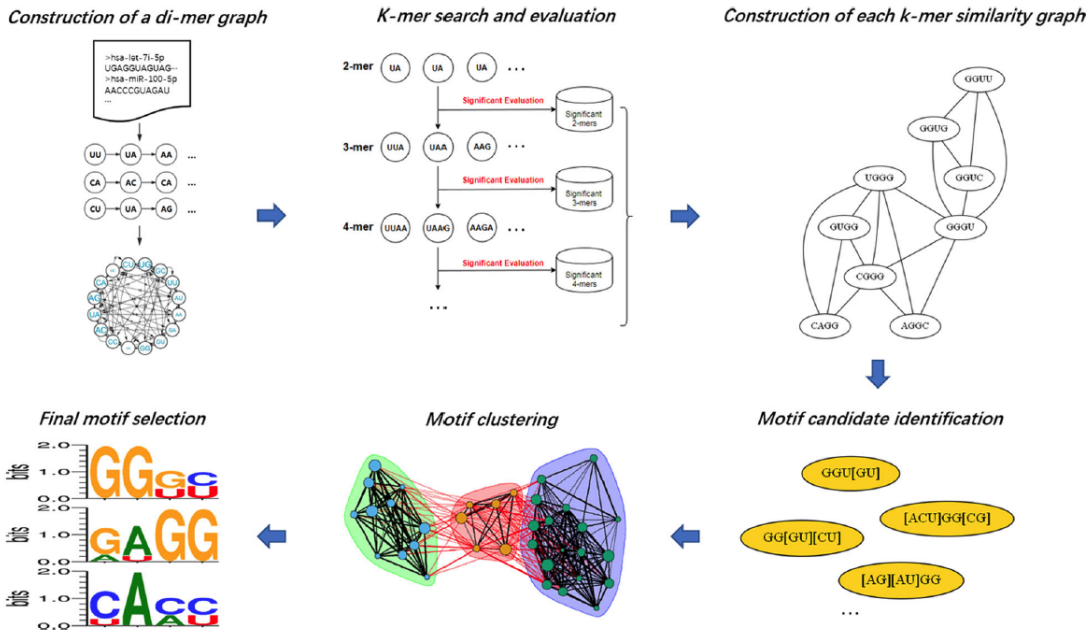

Figure 4.2: Schematic flowchart for motif detection [5]

matrix. Candidate biomarkers were chosen based on microRNAs which were differentially expressed in pancreatic cancer patients and healthy controls.

4.3 Find motifs associated with exosomes (ExoCarta)

In order to explore the molecular determinants in microRNA secretion to further guide the non-invasive biomarker detection in the blood stream, in the second part of this project, we focused on sequence analysis of exosomal microRNAs for motif detection. Particularly, a graph-based motif-finding algorithm previously developed in our lab [5] has been applied for this purpose. The computational procedure for motif detection is shown in Figure 4.2. For details, please refer to the original paper "A systematic approach to RNA-associated motif discovery" [5].

Chapter 5

Results and Discussion

5.1 Visualization of GSE59856 dataset in 2-D and 3-D

Non-invasive screening methods for the early detection of pancreatic cancer is urgently needed. In this study, we investigated the microRNA expression profiles of 100 pancreatic cancer patients and 150 healthy controls (GSE59856, microarray data). Principal component analysis (PCA) was used to visualize sample distribution in two- and three-dimensional scatter plots which reveal separate clusters between the pancreatic cancer and healthy control groups. As shown in Figures 5.1 and 5.2, the overall differentially expressed microRNAs can be used to distinguish cancer samples from healthy controls, except for one outlier which is healthy control.

5.2 Building Machine Learning Models for Cancer Classification

As shown in Table 5.1, Support Vector Machine (SVM), and decision tree classifier (DTC) models performed well. Precision, recall, F1 score, and accuracy reach almost 1 in these cases. On the other hand, linear discriminant anal-

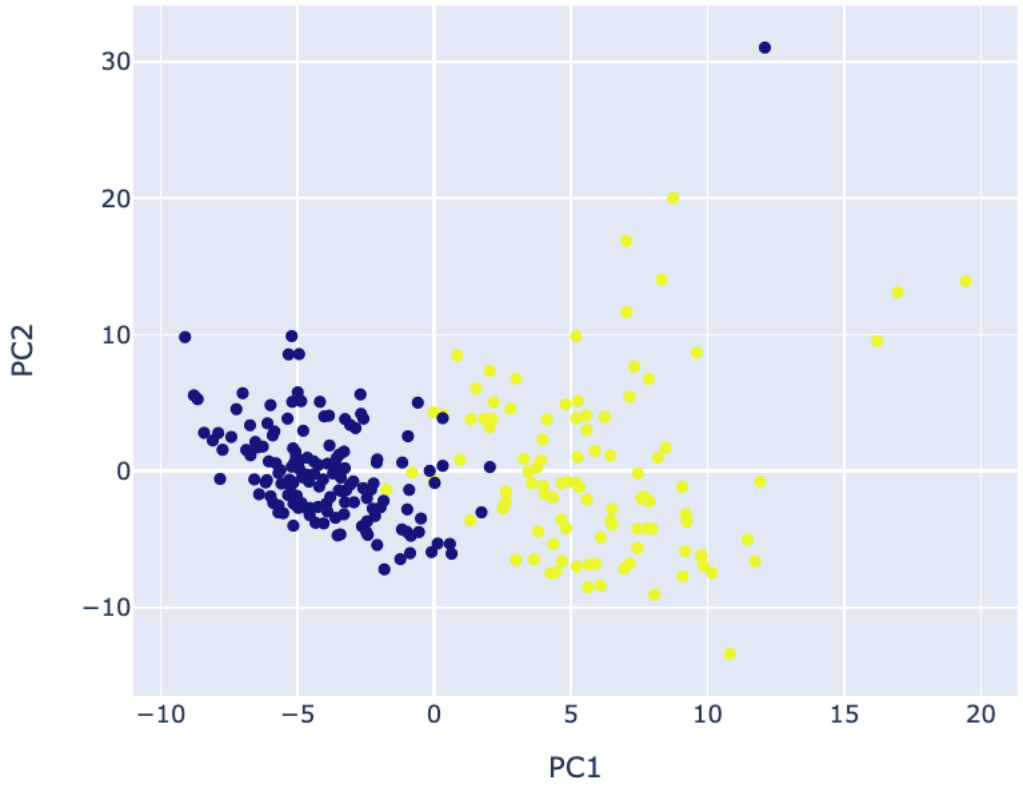

Figure 5.1: Data visualization for GSE59856 in 2-dimension. Yellow: cancer, blue: healthy controls

ysis (LDA), and Gaussian naive bayes (NB), quadratic discriminant analysis (QDA) performed badly. These model could not distinguish cancer samples from normal samples. All standardized evaluation metrics including precision, recall, F1 score, accuracy and area under curve are low. We chose SVM for further analysis because SVM has the highest performance in term of recall. In addition, SVM is well known to be a simple and effective approach.

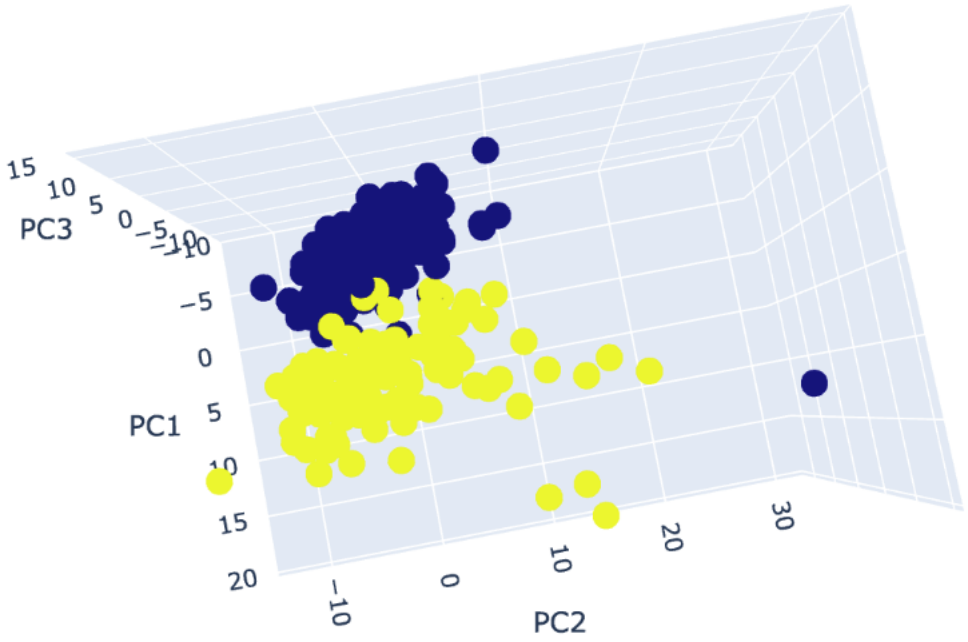

Figure 5.2: Data visualization for GSE59856 in 3-dimension. Yellow: cancer, blue: healthy controls

Model	Recall	Precision	F1 score	Accuracy	AUC
SVM	0.963	0.962	0.959	0.959	0.998
Decision Tree Classifier	0.952	0.964	0.953	0.954	0.952
Linear Discriminant	0.714	0.713	0.717	0.732	0.704
Gaussian Naive Bayes	0.579	0.594	0.577	0.599	0.580
Quadratic Discriminant	0.495	0.448	0.388	0.443	0.569

Table 5.1: Machine Learning Algorithm Comparison

5.3 Candidate biomarkers for detection of pancreatic cancer

We conducted an extensive analysis of publicly available microRNA expression data (GSE59856) to identify candidate exosomal microRNAs associated with pancreatic cancer. A comprehensive list of microRNAs which were differentially expressed between cancer samples and controls is shown in Table A.1. It is believed that cancer cells secrete more exosomes than healthy cells to

cultivate new microenvironments for future metastasis. However, we did not observe this phenomenon in our analysis. Table 5.2 shows the top 20 differentially expressed microRNAs. Among these, 5 microRNAs were up-regulated and 15 were down-regulated in pancreatic cancer. The expression signals of the 4 most significant microRNAs are showed in Figure 5.3. Among those microRNAs, hsa-miR-125a-3p, hsa-miR-6893-5p, and hsa-miR-125b-1-3p were down-regulated in cancer samples while hsa-miR-6075 was up-regulated. The functions of the down-regulated microRNAs could be cancer/tumor suppression. Not having enough these microRNAs could lead to out-of-control cancer/tumor growth. Consistent with the discovery of Kojima et. al. 2015 [11], hsa-miR-125a-3p was found to be the best biomarker for the detection of pancreatic cancer.

We built a SVM model for the classification of pancreatic cancer from healthy control. The effectiveness of the model when combining multiple significant microRNA biomarkers are shown in Table 5.3. When hsa-miR-125a-3p was used as the single biomarker, 28 of 33 (84.85%) pancreatic cancer samples were identified (Table 5.3). When using a combination of the top 5 microRNA biomarkers, the recall rate (true positive divide by the total of true positive and false negative) increased significantly to 96.97%. That means only one out of 33 cancer samples was mistakenly identified as normal sample. With the combination of the top 15 differentially expressed microRNAs, all cancer samples were correctly identified (recall rate = 1). We found that the combination of the full microRNA profile (2555 microRNAs) did not yield the best results (Table 5.3). The reason is that not every microRNA is equally important in distinguishing cancer samples from healthy controls and the presence of ir-

microRNA	logFC	adj.P.Val	P.Value	t	B
hsa-miR-125a-3p	-2.46	3.11E-71	1.22E-74	-26.55	159.23
hsa-miR-6893-5p	-1.35	8.27E-64	6.47E-67	-24.05	141.66
hsa-miR-125b-1-3p	-1.62	8.92E-53	1.05E-55	-20.59	116.12
hsa-miR-6075	1.03	9.87E-49	1.55E-51	19.34	106.61
hsa-miR-4294	-1.02	9.60E-44	1.88E-46	-17.84	95.00
hsa-miR-204-3p	-1.62	3.32E-42	7.79E-45	-17.37	91.31
hsa-miR-1469	0.55	4.38E-42	1.20E-44	17.31	90.88
hsa-miR-575	-1.40	1.14E-39	3.58E-42	-16.59	85.23
hsa-miR-6729-5p	0.25	6.25E-37	2.20E-39	15.78	78.86
hsa-miR-4476	-1.56	5.37E-36	2.10E-38	-15.50	76.62
hsa-miR-6820-5p	-0.72	3.11E-35	1.34E-37	-15.26	74.79
hsa-miR-6765-3p	-1.23	5.03E-32	2.36E-34	-14.32	67.37
hsa-miR-150-3p	-0.98	3.68E-31	1.87E-33	-14.06	65.32
hsa-miR-4792	1.08	1.15E-29	6.28E-32	13.61	61.84
hsa-miR-6836-3p	0.61	2.88E-29	1.69E-31	13.49	60.85
hsa-miR-6799-5p	-0.41	1.14E-28	7.17E-31	-13.30	59.42
hsa-miR-4450	-1.45	8.23E-28	5.47E-30	-13.04	57.41
hsa-miR-4470	-0.95	1.43E-27	1.01E-29	-12.96	56.80
hsa-miR-4419a	-0.98	5.66E-27	4.21E-29	-12.78	55.38
hsa-miR-4530	-0.71	5.87E-27	4.59E-29	-12.77	55.30

Table 5.2: Top 20 circulating microRNA biomarkers that differentiated cancer patients from healthy controls

relevant or partially relevant microRNAs can interfere with cancer diagnosis. To balance the amount of work and the effectiveness of diagnosis methods, the five most significant differentially expressed microRNAs (hsa-miR-125a-3p, hsa-miR-6893-5p, hsa-miR-125b-1-3p, hsa-miR-6075, and hsa-miR-4294) should be used. By using a combination of these five microRNA biomarkers, we were able to distinguish pancreatic cancer patients from those who were healthy, with a diagnostic accuracy of 97.59% and an AUC of 99.70%. For a better performance, a combination of the top 15 microRNAs or top 20 microRNAs should be used for cancer detection.

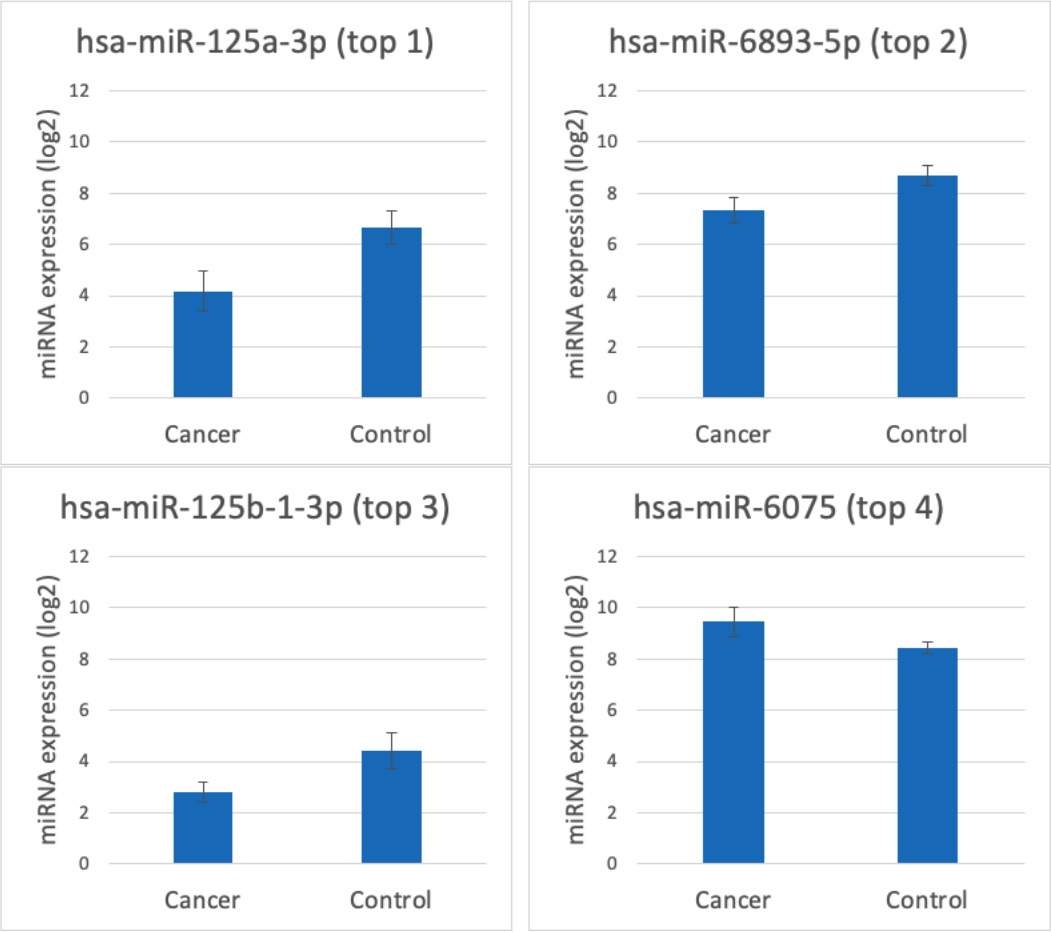

Figure 5.3: Expression signals of the top 4 differentially expressed microRNAs in cancer and healthy control. The error bars are standard deviations.

5.4 microRNA motif sequences associated with exosomes

Motif sequence analysis is important in understanding of microRNA loading mechanism. Previous experimental validation in our laboratory on selected microRNAs in colon cancer cell line showed that mutant of motif sequence shows lowered abundance of exosomal microRNAs compared to the wildtype, which indicates that the motif region is responsible for microRNA loading and packaging [5]. The mechanistic knowledge discovery is the primary goal for motif analysis. In the meantime, although not ideally, we can use motif

Biomarker	Recall	F-1 score	Accuracy	AUC ROC
Top 1 microRNA	0.848	0.918	0.940	0.998
Top 2 microRNAs	0.879	0.935	0.952	0.997
Top 5 microRNAs	0.970	0.970	0.976	0.997
Top 10 microRNAs	0.939	0.969	0.976	0.999
Top 15 microRNAs	1.000	1.000	1.000	1.000
Top 20 microRNAs	1.000	1.000	1.000	1.000
2555 microRNAs	1.000	0.985	0.988	0.999

Table 5.3: The effectiveness of the SVM model for differentiation of cancer samples from healthy controls with different biomarkers.

information to screen for potential secreted microRNA candidates.

Predicted motif sequences highly enriched in exosomal microRNAs are shown in Figure 5.4 and Table 5.4. The higher the information content (IC) and coverage, the more likely that the sequences are true motif sequences.

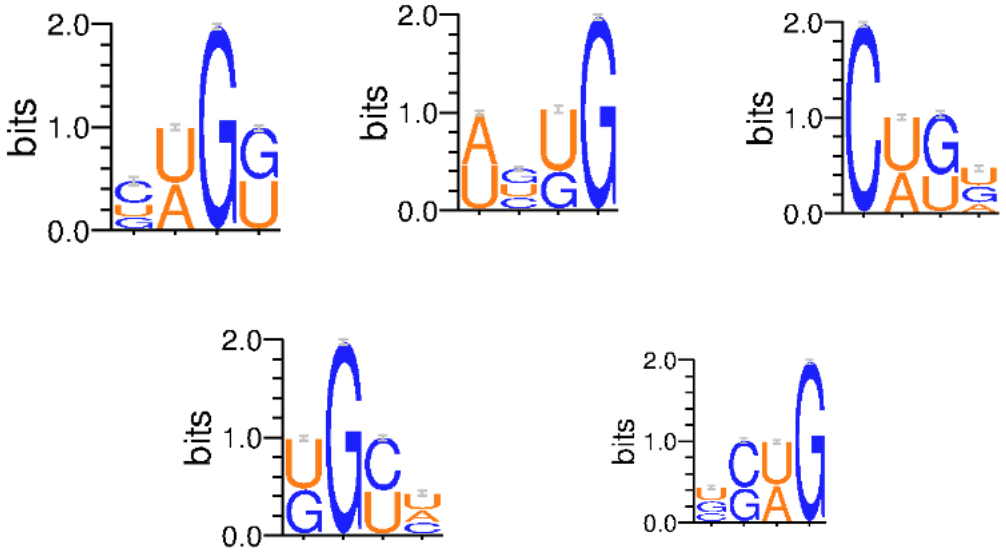

Figure 5.4: Motifs highly enriched in microRNAs associated with exosomes

Motif sequence	**Coverage (in 880)**	**Information Content**	**P value**
[CUG][AU]G[UG]	716	1.34	1.34e-48
[AU][CUG][UG]G	715	1.35	9.14e-47
C[AU][UG][AUG]	714	1.36	1.88e-38
[UG]G[CU][ACU]	713	1.34	2.67e-37
[CUG][CG][AU]G	694	1.35	5.57e-37

Table 5.4: Motif sequences associated with exosomal microRNAs

Chapter 6

Conclusions and Future work

Pancreatic cancer is a deadly disease because it is usually not diagnosed until late stages when treatments are less effective. It is difficult to perform invasive method for the diagnosis of this cancer because of the anatomical position of the pancreas. Therefore, development of non-invasive reliable biomarkers is greatly needed. In this study, we have explored several classification models that can well differentiate cancer from normal based on the expression profiles. Through a series of comparisons and validations, the support vector machine-based classifier has achieved the best performance among all. Using Limma, we identified microRNA candidate biomarkers that were differentially-expressed in cancers versus healthy control. A combination of five significantly-expressed microRNAs (hsa-miR-125a-3p, hsa-miR-6893-5p, hsa-miR-125b-1-3p, hsa-miR-6075, and hsa-miR-4294) achieved the best accuracy and AUC of 97.6% and 99.7%, respectively, which is highly promising. In order to explore the molecular determinants in microRNA secretion to further guide the non-invasive biomarker detection in the blood stream, we focused on sequence analysis of exosomal microRNAs for motif detection. MDS2, a graph-based motif-finding algorithm, was used and the motif [CUG][AU]G[UG] was found highly enriched in microRNAs associated with cancer exosomes. Our research

has further demonstrated that circulating exosomal microRNAs are potential non-invasive biomarkers for early detection of pancreatic cancer. However, more studies including wet laboratory experiments should be conducted to validate the role of these exosomal microRNAs as biomarkers for pancreatic cancer. In addition, research at a large scale should be conducted in this area to discover reliable biomarkers for early detection of pancreatic cancer.

Appendix A

Appendix: Differential Gene Expression

miRNA_ID	miRNA_ID_LIST	logFC	adj.P.Val	P.Value	t	B
MIMAT0004602	hsa-miR-125a-3p	-2.46265	3.11E-71	1.22E-74	-26.5514983	159.2336447
MIMAT0027686	hsa-miR-6893-5p	-1.35424	8.27E-64	6.47E-67	-24.0476507	141.660934
MIMAT0004592	hsa-miR-125b-1-3p	-1.62010667	8.92E-53	1.05E-55	-20.5872088	116.1176164
MIMAT0023700	hsa-miR-6075	1.02857	9.87E-49	1.55E-51	19.34110758	106.6070716
MIMAT0016849	hsa-miR-4294	-1.02188	9.60E-44	1.88E-46	-17.8426717	95.00093014
MIMAT0022693	hsa-miR-204-3p	-1.62281	3.32E-42	7.79E-45	-17.3695698	91.30658797
MIMAT0007347	hsa-miR-1469	0.55154667	4.38E-42	1.20E-44	17.31480761	90.87823522
MIMAT0003240	hsa-miR-575	-1.40042333	1.14E-39	3.58E-42	-16.5939509	85.22833314
MIMAT0027359	hsa-miR-6729-5p	0.24522333	6.25E-37	2.20E-39	15.78344707	78.85929303
MIMAT0019003	hsa-miR-4476	-1.55749	5.37E-36	2.10E-38	-15.4988512	76.62148221
MIMAT0027540	hsa-miR-6820-5p	-0.71875	3.11E-35	1.34E-37	-15.2653905	74.78605461
MIMAT0027431	hsa-miR-6765-3p	-1.22388	5.03E-32	2.36E-34	-14.3205532	67.37029305
MIMAT0004610	hsa-miR-150-3p	-0.97852667	3.68E-31	1.87E-33	-14.058088	65.31671554
MIMAT0019964	hsa-miR-4792	1.07801	1.15E-29	6.28E-32	13.61183115	61.83516557
MIMAT0027575	hsa-miR-6836-3p	0.60857667	2.88E-29	1.69E-31	13.48579308	60.85459911
MIMAT0027498	hsa-miR-6799-5p	-0.40789333	1.14E-28	7.17E-31	-13.3012223	59.4211691
MIMAT0018971	hsa-miR-4450	-1.45381333	8.23E-28	5.47E-30	-13.0409565	57.40548771
MIMAT0018997	hsa-miR-4470	-0.95412333	1.43E-27	1.01E-29	-12.9629568	56.80279842
MIMAT0018931	hsa-miR-4419a	-0.97691	5.66E-27	4.21E-29	-12.778823	55.38279795
MIMAT0019069	hsa-miR-4530	-0.71462667	5.87E-27	4.59E-29	-12.7676884	55.29706019
MIMAT0004804	hsa-miR-615-5p	-0.74105	9.10E-27	7.48E-29	-12.7048968	54.81384506
MIMAT0027574	hsa-miR-6836-5p	0.87229667	1.14E-26	9.86E-29	12.66927799	54.53995997
MIMAT0018976	hsa-miR-4454	-0.78388242	2.31E-26	2.08E-28	-12.5795445	53.80078635
MIMAT0029782	hsa-miR-7641	-1.41821	2.62E-26	2.47E-28	-12.5509869	53.63156223
MIMAT0022945	hsa-miR-1236-5p	-1.02696333	2.75E-26	2.69E-28	-12.5397781	53.54558254
MIMAT0027654	hsa-miR-6877-5p	-0.44874333	5.21E-26	5.30E-28	-12.4519032	52.87211244
MIMAT0004952	hsa-miR-665	0.65543667	5.33E-26	5.64E-28	12.44405472	52.81201439
MIMAT0004748	hsa-miR-423-5p	-0.82222667	6.51E-26	7.14E-28	-12.4134904	52.57805713
MIMAT0000093	hsa-miR-93-5p	-1.07547	7.55E-26	8.57E-28	-12.3897228	52.39621706
MIMAT0031000	hsa-miR-8073	0.69370333	2.70E-25	3.18E-27	12.21982677	51.09880009
MIMAT0018192	hsa-miR-3918	-0.85064859	5.67E-25	6.88E-27	-12.1253292	50.33392442
MIMAT0027660	hsa-miR-6880-5p	-0.75192333	2.60E-24	3.26E-26	-11.9162029	48.79139197
MIMAT0000070	hsa-miR-17-5p	-1.09597333	6.43E-24	8.31E-26	-11.7937119	47.864895
MIMAT0027478	hsa-miR-6789-5p	0.35199	8.09E-24	1.08E-25	11.75961507	47.60746691
MIMAT0005586	hsa-miR-1231	0.53994667	1.05E-23	1.44E-25	11.72185772	47.32264744
MIMAT0026720	hsa-miR-887-5p	-0.75318667	4.17E-23	5.88E-25	-11.5362133	45.92611982
MIMAT0019691	hsa-miR-4634	0.27832667	4.94E-23	7.15E-25	11.5103661	45.73220601
MIMAT0022286	hsa-miR-5585-3p	0.95149667	9.67E-22	1.44E-23	11.11207871	42.76141096
MIMAT0031178	hsa-miR-7975	-0.73372667	1.13E-21	1.72E-23	-11.0881434	42.5839632
MIMAT0000103	hsa-miR-106a-5p	-1.07524667	2.68E-21	4.20E-23	-10.9687763	41.70095295
MIMAT0018179	hsa-miR-3907	-0.90884333	1.16E-20	1.86E-22	-10.768419	40.22632732
MIMAT0019737	hsa-miR-4664-5p	-0.79173899	3.51E-20	5.77E-22	-10.6196998	39.10940092
MIMAT0031180	hsa-miR-7977	-0.69807	3.52E-20	5.92E-22	-10.6120279	39.08213336
MIMAT0023710	hsa-miR-6085	-0.26914	4.48E-20	7.72E-22	-10.576027	38.81962316
MIMAT0019715	hsa-miR-4651	-0.32729	1.29E-19	2.27E-21	-10.4292844	37.75313003
MIMAT0031016	hsa-miR-8089	-0.4462	2.76E-19	4.97E-21	-10.3220415	36.97738248
MIMAT0015070	hsa-miR-3188	0.44487667	3.51E-19	6.46E-21	10.28607223	36.71790883
MIMAT0003228	hsa-miR-564	-1.11032333	3.95E-19	7.42E-21	-10.2671031	36.58121623
MIMAT0019859	hsa-miR-4734	0.27697	4.27E-19	8.19E-21	10.25352186	36.48341149
MIMAT0020601	hsa-miR-1273f	0.70575667	7.01E-19	1.37E-20	10.18251803	35.97293508

Table A.1: Differentially expressed microRNAs between pancreatic cancer samples and healthy controls

Bibliography

[1] S. Ali, H. Dubaybo, R. E. Brand, Sarkar, and F. H. Differential expression of micrornas in tissues and plasma co-exists as a biomarker for pancreatic cancer. *J Cancer Sci Ther*, 7(11):336–346, 2015.

[2] J. M. Arevalillo and H. Navarro. A new method for identifying bivariate differential expression in high dimensional microarray data using quadratic discriminant analysis. *BMC Bioinformatics*, 2011.

[3] N. Dilsiz. Role of exosomes and exosomal micrornas in cancer. *Future Sci OA*, 6(4), 2020.

[4] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky. mirdeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2012.

[5] T. Gao, J. Shu, and J. Cui. A systematic approach to rna-associated motif discovery. *BMC Genomics volume*, 19(146), 2018.

[6] T. Goto, M. Fujiya, H. Konishi, J. Sasajima, S. Fujibayashi, A. Hayashi, T. Utsumi, H. Sato, T. Iwama, M. Ijiri, A. Sakatani, K. Tanaka, Y. Nomura, N. Ueno, S. Kashima, K. Moriichi, Y. Mizukami, Y. Kohgo, and T. Okumura. An elevated expression of serum exosomal microrna-191,

21, 451a of pancreatic neoplasm is considered to be efficient diagnostic marker. *BMC Cancer*, 18(116), 2018.

[7] M. Hassan and R. Kotagiri. A new approach to enhance the performance of decision tree for classifying gene expression data. *BMC Proc.*, 7(S3), 2013.

[8] D. Huang, Y. Quan, M. He, and B. Zhou. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental Clinical Cancer Research*, 28(149), 2009.

[9] N. C. Institute. Pancreatic cancer treatment. 2018.

[10] K. Kim, D. Yoo, H. S. Lee, K. J. Lee, S. B. Park, C. Kim, J. H. Jo, D. E. Jung, and S. Y. Song. Identification of potential biomarkers for diagnosis of pancreatic and biliary tract cancers by sequencing of serum micrornas. *BMC Medical Genomics*, (62), 2019.

[11] M. Kojima, H. Sudo, J. Kawauchi, S. Takizawa, S. Kondou, H. Nobumasa, and A. Ochiai. Microrna markers for the diagnosis of pancreatic and biliary-tract cancers. *PLoS One*, 10(2), 2015.

[12] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 2019.

[13] X. Lai, M. Wang, S. D. McElyea, S. Sherman, M. House, and M. Korc. A microrna signature in circulating exosomes is superior to exosomal glypican-1 levels for diagnosing pancreatic cancer. *Cancer Lett*, 393.

[14] B. Lan, S. Zeng, R. Grützmann, and C. Pilarsky. The role of exosomes in pancreatic cancer. *Int J Mol Sci.*, 20(8):1–14, 2019.

[15] J. Lee, H. S. Lee, S. B. Park, C. Kim, K. Kim, D. E. Jung, and S. Y. Song. Identification of circulating serum mirnas as novel biomarkers in pancreatic cancer using a penalized algorithm. *Int J Mol Sci*, 22(3):1007, 2021.

[16] L.-A. Macfarlane and P. Murphy. Microrna: Biogenesis, function and role in cancer. *Current genomics*, 11(7):537–561, 2010.

[17] T. Machida, T. Tomofuji, T. Maruyama, T. Yoneda, D. Ekuni, T. Azuma, H. Miyai, H. Mizuno, H. Kato, K. Tsutsumi, D. Uchida, A. Takaki, H. Okada, and M. Morita. mir-1246 and mir-4644 in salivary exosome as potential biomarkers for pancreatobiliary tract cancer. *Oncol Rep*, 36(4):2375–81, 2016.

[18] B. Madhavan, S. Yue, U. Galli, S. Rana, W. Gross, M. Müller, N. A. Giese, H. Kalthoff, T. Becker, M. W. Büchler, and M. Zöller. Combined evaluation of a panel of protein and mirna serum-exosome biomarkers for pancreatic cancer diagnosis increases sensitivity and specificity. *Int J Cancer*, 136(11):2616–27, 2015.

[19] S. Nakamura, Y. Sadakari, T. Ohtsuka, T. Okayama, Y. Nakashima, Y. Gotoh, K. Saeki, Y. Mori, K. Nakata, Y. Miyasaka, H. Onishi, Y. Oda, M. Goggins, and M. Nakamura. Pancreatic juice exosomal micrornas as biomarkers for detection of pancreatic ductal adenocarcinoma. *Ann Surg Oncol*, 26(7), 2019.

[20] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng. Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, 9:402, 2018.

[21] R. Que, G. Ding, J. Chen, and L. Cao. Analysis of serum exosomal micrornas and clinicopathologic features of patients with pancreatic adenocarcinoma. *World J Surg Oncol.*, 2013.

[22] K. W. . K. T. . S. M. . M. K. . M. S. . F. M. . K. S. . Sachiyo Kawamura 1, Hisae Iinuma 1. Exosome-encapsulated microrna-4525, microrna-451a and microrna-21 in portal vein blood is a high-sensitive liquid biomarker for the selection of high-risk pancreatic ductal adenocarcinoma patients. *J Hepatobiliary Pancreat Sci.*, 26(2):63–72, 2019.

[23] K. Takahasi, H. Iinuma, K. Wada, S. Minezaki, S. Kawamura, M. Kainuma, Y. Ikeda, M. Shibuya, F. Miura, and K. Sano. Usefulness of exosome-encapsulated microrna-451a as a minimally invasive biomarker for prediction of recurrence and prognosis in pancreatic ductal adenocarcinoma. *J Hepatobiliary Pancreat Sci*, 25(2):155–161, 2018.

[24] D. Taller, K. Richards, Z. Slouka, S. Senapati, R. Hill, D. B. Go, and H.-C. Chang. On-chip surface acoustic wave lysis and ion-exchange nanomembrane detection of exosomal rna for pancreatic cancer study and diagnosis. *Lab Chip*, 15(7):1656–1666, 2015.

[25] Wikipedia. Naive bayes classifier.

[26] Y. Xu, X. Xu, A. Williams, and W. Ding. The role of exosomal micrornas in pancreatic cancer. *Stem Cell Investigation*, 7, 2020.

[27] Y.-F. Xu, B. N. Hannafon, U. Khatri, A. Gin, and W.-Q. Ding. The origin of exosomal mir-1246 in human cancer cells. *RNA Biol*, 16(6):770–784, 2019.

[28] Y.-F. Xu, B. N. Hannafon, Y. D. Zhao, R. G. Postier, and W.-Q. Ding. Plasma exosome mir-196a and mir-1246 are potential indicators of localized pancreatic cancer. *Oncotarget*, 8(44):77028–77040, 2017.

[29] Z. R. Yang. Biological applications of support vector machines. *Brief Bioinform.*, 5(4):328–338, 2004.